Supplementary Figure Legends

Fig. S1: Description of the TCGA data. A) Number of samples from each tumor type among discovery and validation cohorts. B) Number of samples in the discovery and validation cohorts with data for each somatic phenotype. C) Self-reported ancestry for 6,908 samples from TCGA meta-data. TCGA 4 letter codes: BRCA:Breast Invasive Carcinoma, GBM:Glioblastoma Multiforme, OV:Ovarian Serous Cystadenocarcinoma, LUAD:Lung Adenocarcinoma, UCEC:Uterine Corpus Endometrial Carcinoma, KIRC:Kidney Renal Clear Cell Carcinoma, HNSC:Head and Neck Squamous Cell Carcinoma, LGG:Brain Lower Grade Glioma, THCA:Thyroid Carcinoma, LUSC:Lung Squamous Cell Carcinoma, PRAD:Prostate Adenocarcinoma, STAD:Stomach Adenocarcinoma, SKCM:Skin Cutaneous Melanoma, COAD:Colon Adenocarcinoma, BLCA:Bladder Urothelial Carcinoma, LIHC:Liver Hepatocellular Carcinoma, CESC:Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma, KIRP:Kidney Renal Papillary Cell Carcinoma, LAML:Acute Myeloid Leukemia, PAAD:Pancreatic Adenocarcinoma, READ:Rectum Adenocarcinoma, KICH:Kidney Chromophobe

Fig. S2: Flow chart describing data preprocessing, association testing and validation testing stages of the study. Numbers of markers evaluated and number of samples used are detailed for each stage.

Fig. S3: QQ-plots and genomic inflation (G.I.) estimates for germline associations with tumor site of origin. QQ-plots display observed versus expected p-values given the number of statistical tests performed for each tumor type. A red diagonal represents the expected distribution. Points to the left of the diagonal represent associations that are more significant than expected. Genomic inflation estimates were used to correct p-values for selecting candidate associations in the discovery screen.

Fig. S4: Odds ratios for published cancer-associated markers from the NHGRI GWAS Catalog were compared to odds ratios for the same markers when testing for associations with tumor site of origin. Associations with tumor site of origin were evaluated specifically in the tumor type for which the marker was originally reported. Odds ratios are compared for associations with A) breast cancer, B) colon cancer, C) pancreatic cancer and D) prostate cancer. Dark blue lines represent a linear regression of ORs in TCGA onto published ORs.

Fig. S5: Power analysis and comparison of effect sizes between discovery and validation phases A) Power curves describing estimated power to detect associations dependent on sample size and effect size assuming a variant minor allele frequency of 1% (top). Colored curves show power for different effect sizes. A histogram shows the number of samples with a somatic alteration for each of 138 cancer genes (bottom). Genes are colored according to role as a tumor suppressor (blue) or oncogene (red). B) Odds ratios for validated cancer-gene associated markers compared between the discovery and validation screens. Odds ratios show strong positive correlation ($r=0.043$; $P < 0.005$).

Fig. S6: QQ-plots for germline associations with somatic status of cancer genes with validated associations. A red diagonal represents the expected distribution. Points to the left of the diagonal represent associations that are more significant than expected. Genomic inflation estimates were 1 for all genes displayed. QQ-plots with two colored lines show the distribution of p-values for somatic-germline associations when significant associations were detected for multiple types of somatic perturbations (i.e. statistically significant associations were detected when by grouping patients with mutation or with CNV, or considering the union of mutation and CNV). Groupings are labeled with the color of the corresponding line.

Fig. S7: Western blots. These blots demonstrate A) the reproducibility of experimental knockdown of PTEN using siRNAs at 4 concentrations, and B) the reproducibility of STK11 knockout by CRISPR-CAS9.

Supplementary Table Legends

Table S1: Validated associations identified between germline markers and tumor site of origin. Each row describes a single association.

Table S2: Results from re-testing published cancer GWAS associations from the NHGRI GWAS Catalog in the corresponding tumor type in the TCGA. Differential expression for each associated marker was evaluated separately in relevant tumor types, and p-values were corrected for the number of genes as well as the number of tumor types tested.

Table S3: Validated associations identified between germline markers and alteration status of cancer genes. Each row describes a single association.

Table S4: TCGA samples used for association analysis.

Table S5: TCGA samples used for alternative splicing analysis.